Inside Market Data

March 2014

waterstechnology.com/imd

LATENCY

SPECIAL REPORT

Sponsored by:

AZUL SYSTEMS

Quincy Data LLC

THOMSON REUTERS

Why not choose the only Java platform that starts fast and stays fast?

From the opening bell, Zing® delivers consistent, low latency Java performance for capital markets.

- Predictable, proven jitter-free operation
- New ReadyNow! Technology that solves Java "warm-up" problems
- No code changes or special libraries required
- Broad partner network for trading/financial services
- Meets Java SE 6 and SE 7 standards
- Industry-leading Java support

Learn how Zing can give you a competitive edge
www.azulsystems.com/low-latency-java

Azul Zing®

The New Java Performance Standard

Azul Systems, 1173 Borregas Ave., Sunnyvale, CA 94089-1306 USA | + 1.650.230.6500 | www.azulsystems.com

AZUL SYSTEMS®

LETTER FROM THE EDITOR

Welcome to the Latency Games

Whether you consider it a modern classic or not, *The Hunger Games* by Suzanne Collins teaches a valuable lesson about competition. The moral of the story is akin to Aesop's Tortoise and the Hare fable, in which an over-confident bunny takes a nap mid-race and loses to a slow-moving tortoise. You don't have to be the fastest to win the race.

When I joined *Inside Market Data* three years ago, the buzz around latency was at fever pitch and the race to zero was dominated by firms with fat wallets who could afford to spend millions to save a few microseconds.

Since then, the fervor has cooled, but as Bill Ruvo, global business head of real-time feeds at Thomson Reuters, notes in a sponsored statement *Latency: Dead or Just Misunderstood?* on page 12 of this report, the industry now operates all along the latency spectrum. Many firms have realized that they can be fast if not the fastest, and still be successful, while the commoditization of low-latency technology means that most can operate at the edge, if not the bleeding edge.

Indeed, some might argue that the latency race has evolved into something much more strategic; the Latency Games. And just like Katniss Everdeen in *The Hunger Games*, firms are doing everything to make sure the odds are in their favor.

Take for example, latency monitoring, which has moved beyond the simple measurement of latency to actively predicting where latency issues might occur in the future. The most successful players are those whose networks are most reliable, whose packets are most consistent, and whose feeds are jitter-free. Meanwhile, in the network space, firms are starting to experiment with microwave technology, recognizing that a hybrid of fiber and microwave might be better suited to their needs.

In this new world, the winners might not be the fastest, but they play the game the best. So welcome, welcome, welcome to the Latency Games! May the odds be ever in your favor. ■

Faye Kilburn
US Reporter, *Inside Market Data*

Inside Market Data

Max Bowie, **Editor**
Tel: +1 646 490 3966
max.bowie@incisivemedia.com

Faye Kilburn, **US Reporter**
Tel: +1 646 490 3967
faye.kilburn@incisivemedia.com

Giulia Lasagni, **European Reporter**
Tel: +44 (0)20 7316 9143
giulia.lasagni@incisivemedia.com

Lee Hartt, **Group Publishing Director**
Tel: +44 (0)20 7316 9443
lee.hartt@incisivemedia.com

Jo Webb, **Global Commercial Director**
Tel: +44 (0)20 7316 9474
jo.webb@incisivemedia.com

Bene Archbold, **US Commercial Manager**
Tel: +1 646 736 1892
bene.archbold@incisivemedia.com

Elina Patler, **Head of Editorial Operations**
Claire Light, **Senior Marketing Manager**
Natasha Gordon-Douglas, **Marketing Manager**

Incisive Media
55 Broad Street, 22nd Floor
New York, NY 10004
Tel: +1 646 736 1888

Incisive Media
Haymarket House
28–29 Haymarket
London SW1Y 4RX
tel: +44 (0)20 7316 9000
fax: +44 (0)20 7930 2238

Incisive Media
14th Floor (Unit 1401-3), Devon House, Taikoo Place
979 King's Road Quarry Bay, Hong Kong
Tel: +852 3411 4900

Subscription Sales
Monazer Rashid Tel: +44 (0)20 7968 4506
waters.subscriptions@incisivemedia.com

Incisive Media Customer Services
E-mail: customerservices@incisivemedia.com
Tel (UK): 0870 787 6822
Tel (International): +44 (0)1858 438421

An incisivemedia publication

NEWS ROUNDUP

CBOE Lays Out 12-Month Latency Cut Plan

The Chicago Board Options Exchange is in the middle of a project to reduce data and transaction latency across its trading systems to less than 100 microseconds, building on existing latency reductions achieved as a result of moving its primary matching engines from Chicago to the New York metro area just over a year ago.

Mark Novak, vice president and chief technology officer of systems development at CBOE, says the migration to Equinix's NY4 datacenter in Secaucus, NJ on Dec. 3, 2012 reduced roundtrip latency for New York-based trading participants from around 12 milliseconds to 1 millisecond or less.

Along with the move, the exchange took the opportunity to upgrade its internal infrastructure to 10 Gigabits per second, and upgraded its hardware from servers leveraging Intel's Westmere microprocessors to servers running Intel's newer Sandy Bridge family of chips. As a result of these upgrades and other software changes, CBOE slashed internal transaction latency—the quote-to-trade time from a firm submitting an order to that trade appearing on CBOE's datafeed—from around 600 microseconds to around 200 microseconds.

Now, CBOE is working on initiatives to further reduce that latency figure to around 100 microseconds by optimizing its server processing paths and message transport between different levels of the architecture, Novak says.

"To get below 100 microseconds, we look at other things, like removing a tier from our architecture to create a more direct path from our participants to our trading servers—for example, by using an FPGA switch-based approach to routing messages, removing part of the interface tier," he says. "The functionality won't change… but every tier in the architecture adds latency, so we can cut down that transport layer."

Novak calls the latency reduction efforts "an ongoing process" with an evolving goal, which is expected to take the next 12 months or more to complete. ■

TSE Network Clients Get Corvil Latency Measurement

The Tokyo Stock Exchange (TSE) will offer Dublin-based technology vendor Corvil's CorvilClear latency performance management solution to trading firms connected to its Arrownet network on an as-a-service basis, to provide them with visibility into latency and trade processing latency between themselves and the exchange.

TSE has used technology from Corvil internally for latency management of its Arrowhead and Tdex+ trading systems and Arrownet network since 2012, which officials say has improved the performance of the exchange's order processing and market data distribution. ■

Morningstar Australia Ticker Plant to Cut Latency by 100ms

Chicago-based data and investment research provider Morningstar has set up a ticker plant in its Sydney, Australia datacenter to provide lower-latency content via its datafeeds to local trading firms.

Officials say Morningstar's new ticker plant—which joins other local ticker plants throughout the Asia-Pacific region in Singapore, Hong Kong, Shanghai and Tokyo—will reduce latency by up to 100 milliseconds as well as reducing firms' infrastructure costs compared to consolidated datafeeds that collect data at a regional rather than market-specific level. ■

Perseus Bakes Low-Latency London Microwave Network

New York-based fiber and microwave network provider Perseus Telecom has built a series of wireless microwave networks connecting key datacenters in and around London to provide market makers and latency-sensitive traders with ultra-low latency connectivity to support market making and arbitrage trading in the equities, derivatives and foreign exchange markets.

Perseus went live with four new microwave routes—between Equinix's LD4 datacenter in Slough and the London Hosting Center in London's Docklands district; between LD4 and NYSE Euronext's Liquidity Center datacenter in Basildon, Essex; between LD4 and the London Stock Exchange's datacenter; and between NYSE Basildon and the LSE—in the second week of January, and already has trading firm clients using the networks, says Perseus chief executive Jock Percy.

The vendor expects each wavelength to yield capacity of between 120 and 150 Megabits per second (Mbps), and will sell segregated circuits of bandwidth in 10Mbps increments.

Percy says clients will typically use the microwave circuits for subscribing to relatively small quantities of market data and for sending orders—though Perseus has also offered fiber networks on the same routes since 2010 for extra resiliency. However, he says the microwave networks will reduce latency over the same routes by up to 30 percent. Firms using the microwave networks—which have limited bandwidth compared to fiber or millimeter-wave wireless—can mitigate the limited bandwidth by fine-tuning their strategies and trading systems to use less bandwidth.

Traders can use the networks to arbitrage equities listed on multiple marketplaces or derivatives against their underlying equities, or to arbitrage between Thomson Reuters' FX market in the LHC datacenter and Icap's EBS FX market located in LD4. ■

Maple Leaves HFT Co-Lo for SpryWare Cloud

A proprietary trading group at Maple Securities USA, the US subsidiary of Canada's Maple Financial Group, is finalizing a migration from direct exchange feeds and an in-house co-location operation to a managed market data infrastructure run by Chicago-based low-latency data and technology vendor SpryWare in its Proximity Cloud hosted data service.

The cutover process will be complete when Maple removes its trading servers from Nasdaq's Carteret, NJ datacenter, having already moved its technology to the firm's new office in Hoboken, NJ, after going live on the SpryWare service around the start of this year.

"We've been co-located and using raw feeds from market centers for years, but that was becoming overwhelming," in terms of the volumes of data, number of venues required to connect to, and the amount of maintenance to update exchange data formats and constantly build order books from raw data, says Kevin Keane, head of trading for Maple's systems trading group. "We've backed away from high-frequency trading—now our trading frequency is going out to holding periods of a day or a week—so that means we don't need to be co-located or worry about using microwaves or lasers [for low-latency data distribution]. And if we're not in the game of trying to be first, then letting someone else do all that for us is great."

Instead, Maple now hosts its trading systems in SpryWare's Proximity Cloud, where they subscribe to a feed of data on around 2,000 US equities sourced from Pico Quantitative Trading via its partnership with SpryWare. Maple signed up with SpryWare in September, and began running SpryWare's service in parallel with its existing co-lo site last year, then cut over to SpryWare as its primary data source in late December or early January, Keane says. ■

xCelor Debuts Single-Digit Nanosecond Data Switches

Chicago-based low-latency switching and data technology vendor xCelor has begun rolling out its family of xPort Layer 1 network switches, which replicate data from feeds without performing other processing tasks, to minimize port-to-port latency to between 2 and 4 nanoseconds.

Pricing for the devices is tiered based on the functionality and number of ports supported, ranging from $13,000 for the XPR with 16 ports to $23,000 for the XPM with 48 ports. ■

Spread Connects NYSE Mahwah Datacenter to Chicago

Ridgeland, Miss.-based low-latency fiber network operator Spread Networks has set up a point of presence for its wavelength data network between New York and Chicago at NYSE Euronext's datacenter at Mahwah, NJ, creating roundtrip latency of around 14 milliseconds between NYSE's markets hosted at the facility and key markets located in Chicago.

NYSE clients will be able to access Spread Networks' wavelengths via NYSE's SFTI network services business. ■

BATS–Direct Edge to Equalize NY4, NY5 Latency, Post-Equinix Move

BATS Global Markets will ensure that trading members of its BATS and Direct Edge exchanges will experience the same latency whether they co-locate in datacenter provider Equinix's Secaucus, NJ NY4 or NY5 facilities when the US equities and options exchange operator consolidates its matching engine platforms on BATS technology in the new NY5 datacenter next year, as part of a post-merger integration process that includes choosing Equinix as the combined exchange group's primary datacenter provider.

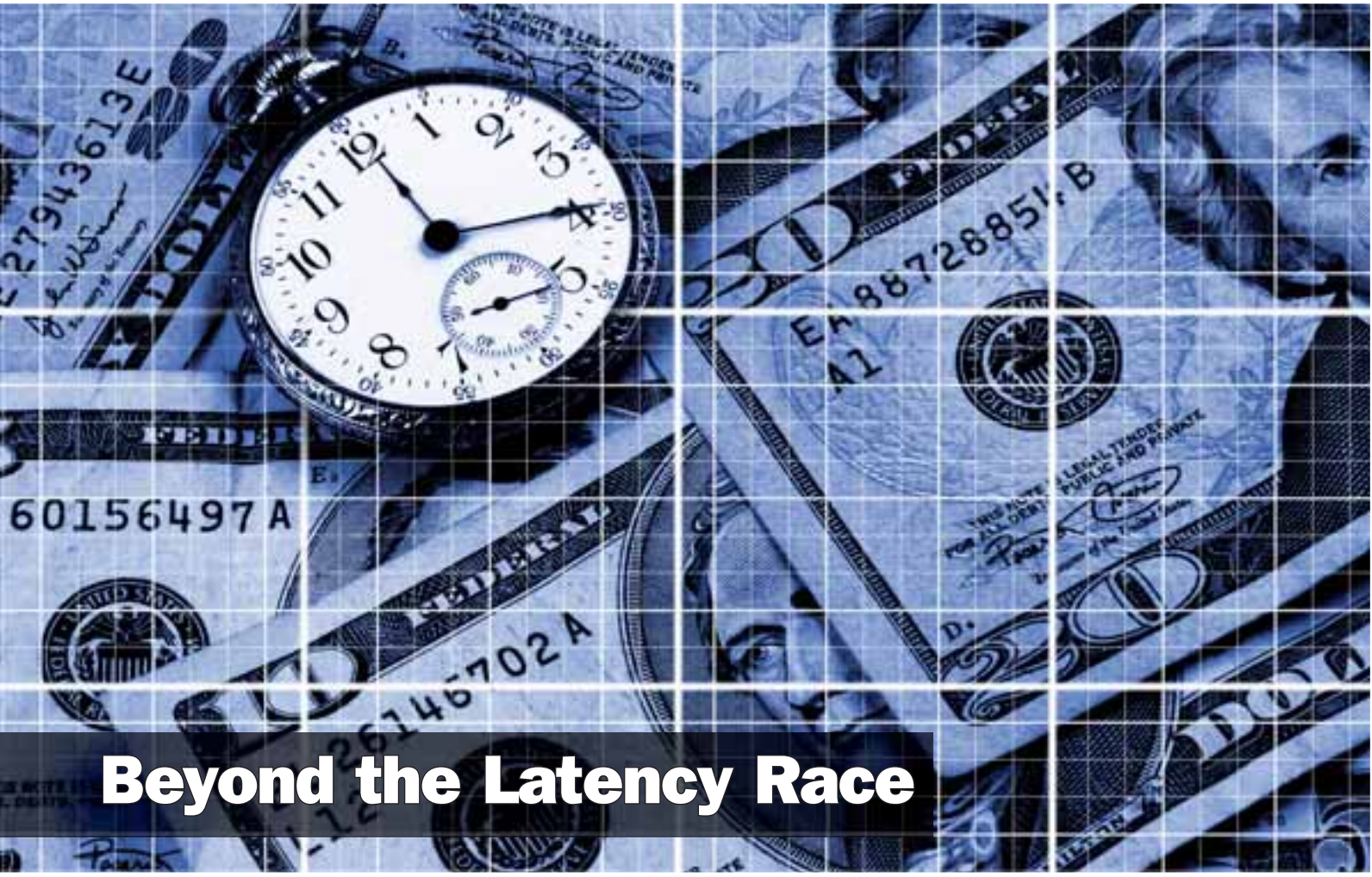
In January 2015, BATS will migrate the matching engines for Direct Edge's EDGA and EDGX platforms currently located in Equinix's NY4 datacenter in Secaucus onto BATS technology hosted in Equinix NY5—also located in Secaucus, as part of the same datacenter campus as NY4 and Equinix's NY2 datacenter. Then in the second quarter of next year, BATS will move its BZX, BYX and BATS Options markets from rival datacenter provider CenturyLink Technology Solutions' (formerly Savvis) Weehawken datacenter to NY5.

To ensure that clients moving into NY5 do not obtain a latency advantage over those with existing infrastructure investments in NY4 and who choose to remain in that facility, BATS will introduce a tiny delay to ensure trading firms in both datacenters receive its market data at the same time.

"We plan to engineer latency so it is equal for anyone connecting from NY4 or NY5," says BATS global chief information officer Chris Isaacson, adding that the length of cable that would run between the two datacenters within the campus is less than half a mile, so the latency introduced would only amount to single-digit microseconds. "NY4 is a very heavily-utilized datacenter. We want to preserve that as much as possible, but we recognize that there is not enough room for everyone to move there."

As an added incentive for clients to move to NY5, BATS negotiated a discount program with Equinix for firms that co-locate in the datacenter to connect to BATS and Direct Edge. Isaacson declines to specify details of the discount. ■

ROUNDTABLE



Beyond the Latency Race

Despite the commoditization of low-latency technology, many trading firms have dropped out of the latency race; either they do not have the capital to compete with the fastest or the returns are insufficient to justify the investment. So is the latency race over for good, or just entering a new phase of complexity?

IMD: Where do you see your firm's place in the latency race? What aspects of market data and trading infrastructure are most important to monitor, and where can the most savings be made?

Stéphane Tyc, co- founder, Quincy Data: At Quincy Data, we are leading the effort to bring microwave market data into the mainstream. We stand for transparency. We publish our latencies and our prices on our website and we even indicate when we believe we have the fastest path. Quincy already offers the fastest and broadest ultra-low latency data service and we are delivering improvements rapidly.

While Quincy often approaches the speed of light in air, absolute latency is not the most important factor to track. It is even more important to monitor the relative latency of different participants on any given route and the ability for all firms willing to pay a reasonable price to have access to that route. The very fabric of the market is deeply influenced by the relationship of relative latency and accessibility. When the fastest link is offered as an affordable service like Quincy's—rather than dominated by a single participant—competition is enhanced, the playing field becomes more level and the market is healthier and fairer.

Bill Ruvo, global business head of real-time feeds, Enterprise capabilities, Thomson Reuters: As an established market data and technology vendor, we view our position as an essential client partner, delivering differing levels of service to meet the objectives of varied trading strategies. We understand that the ultra-latency sensitive market segment will not consider vendor services; however, the universe outside this segment is large, with diverse needs. To become the partner of choice, it's essential to provide a suite of services that meets a broader

THOMSON REUTERS

Bill Ruvo
Global Business Head, Real-Time Feeds,
Enterprise Capabilities
Thomson Reuters
Tel: +1 646 223 4566
Email: bill.ruvo@thomsonreuters.com
Web: www.thomsonreuters.com

"In communications, the two most interesting technologies are millimeter wave and hollow-core fiber. Millimeter wave is much like microwave but allows for more efficient use of the available RF spectrum, which means that more bandwidth can be delivered than across a comparable microwave link."

Bill Ruvo, global business head of real-time feeds, Enterprise capabilities, Thomson Reuters

latency profile with consistent standards around APIs, data models, symbology and coverage, and which maps to the different workflows and business activities under consistent licensing fees. The importance of monitoring the data and the infrastructure again depends on the particular strategy. End-to-end monitoring is proportionately critical to latency sensitivity. Latency measurement is particularly critical in co-location scenarios and frankly, firms engaged in high frequency trading must commit the dollars necessary to monitor performance to an exacting level. In such cases, it's typically not about cost savings but more about alpha generation.

Gil Tene, vice president of technology and chief technical officer, Azul Systems: At Azul, we focus purely on Java, and we find this is used quite a bit in low-latency applications. We see our position as enabling Java to be used successfully in this space, erasing the historical limitations of Java, and creating a better environment for low-latency applications.

We focus on the consistency race; not just speed. Being fast for only part of the time just doesn't cut it. Being fast consistently and being able to predict consistency is just as important as being fast on its own.

Every step in the execution path is critical to monitor, in order to understand which parts of the execution systems need work. This is valuable not just for performance monitoring, but also to help improve those systems. It's important to be able to answer where something happened and what changed, as opposed to just knowing that something happened—especially when there are multiple parties involved.

When people talk about savings, they are usually asking where they can shave off more latency. But we actually look for where we can make latency more consistent. Improving the 99th percentile consistency increases profit and reduces risk by more than improving the median latency.

Mehmet Yanilmaz, partner at Myra Trading: Myra Trading doesn't treat latency as a race to be won, but rather as a key ingredient of trading strategies that needs to be managed holistically across the whole trading ecosystem. The most important saving can be made by concurrent tracking of variations in the performances of trading platforms' matching engines, throughput rates of data transmission media, and coordinating these measures with time stamping rates within the firm's internal ticker plants and algorithms' reaction times to market opportunities.

Peter Nabicht, consultant and senior advisor, Modern Markets Initiative: Firms shouldn't view lowering their latency as a race to be won or lost. Low latency is a valuable tool when properly applied, but should only be applied to the trading strategies that benefit from the latency improvements.

No matter what the latency requirements of your trade are, it makes sense to monitor closely all aspects of market data and trading infrastructure because it is crucial that the data on which you are making your decisions is consistent and reliable.

I think most savings can be found in right-fitting your infrastructure to your needs. The low-latency investments you don't make because your strategy doesn't require them are the best cost savings.

IMD: Is latency monitoring more important than simply being fastest? How is latency monitoring technology being used more broadly to understand different aspects of data flow?

Yanilmaz: Yes, latency monitoring is more important than simply being the fastest. Integrating the latency portfolio resulting from the holistic monitoring approach described above in real time with the firm's portfolio of trading algorithms is the most important factor of a successful latency management strategy.

Tene: Latency monitoring is critical for understanding the behavior of latency. What's more critical than simply being fast is knowing how latency behaves, especially when you're doing millions of things, each of which has a different latency level. Monitoring is responsible for telling us that behavior. Without it, there is no way to shape latency to be better. I've been toying with the term "latency volatility"—the idea that this is more important today than absolute speed. And the ability to monitor is very important to understanding how that volatility behaves.

ROUNDTABLE

Quincy Data LLC

Stéphane Tyc
Co-Founder
Quincy Data
Tel: +33 (0) 9 72 33 76 18 (France)
+1 201 884-1114 (US)
Email: stephane.tyc@quincy-data.com
Web: www.quincy-data.com

"For risk takers and managers, the important question is who they can rely on to consistently deliver—in terms of latency, reliability and data breadth—and what do they need to run their businesses effectively today and at each step toward zero latency? We have a very strong track record of delivering for clients. In the end, firms no longer need to compete on speed."

Stéphane Tyc, co-founder, Quincy Data

Ruvo: Latency monitoring is traditionally associated with being fastest, but in these times of changing markets and increased regulation, it plays a bigger role ensuring accuracy and best execution. For simple high-frequency trading strategies, latency carries significant importance to the success of the strategy, and therefore must be continuously measured and constantly monitored. For more complex strategies, latency is not quite so critical to success, but still must be measured and monitored to ensure such things as best execution and regulatory compliance.

Nabicht: I think latency monitoring is rarely important to the fastest. There is no need to make the investment in being the fastest if your trading strategy doesn't gain enough profit from the improvements to offset the cost. You simply need to be fast enough to properly execute your strategy and to manage your risk.

That said, monitoring your latency, data processing and any queues is very important. You need to know if your live performance matches the latency characteristics your strategy expects. If your strategy runs higher risks or is no longer profitable at certain speeds, it is good to know so you can take appropriate actions.

Tyc: Latency monitoring is important, but it is even more important to design your networks and applications to avoid jitter. If you are careful enough, it is possible to reduce jitter in your system overall. Quincy's data service all but eliminates jitter by doing conflation instead of buffering when it is necessary. Understanding the sources of latency on the exchange side and the different data channels and offerings is key.

IMD: What are the advantages and limitations to microwave technology? Will microwave technology ever become mainstream, or will the high cost and limitations prohibit widespread adoption?

Tyc: Microwave is already becoming mainstream. Faster decision cycles combined with simple arithmetic are largely driving this process. Most firms' decision cycles are under 10 milliseconds (ms) today—and declining rapidly—versus 1000ms a decade ago. Microwave's advantage over fiber is a physical constant, varying only by the length of the route. As soon as your decision cycle time is less than the difference between microwave and fiber, you have to move to microwave; otherwise you base your decisions on stale data. It is as simple as that.

Another reason microwave is moving into the mainstream is that the costs associated with microwave data are already competitive compared to the advantages they provide. Quincy Data's service is designed to be extremely affordable. We try to make it very simple for firms to access the fastest data, so we post our price list on our website, we offer three-month contracts and a discount for small firms. Fast, flexible, affordable and transparent seems like a mainstream product to us.

Ruvo: Microwave is a great way to reduce latency between two points. The advantages begin with the fact that the refractive index of light in air is close to one, as opposed to fiber which is closer to 1.5, which means that microwave in air propagates in two thirds the time that light propagates in fiber. Add to that the fact that light travels in a straight line in air, while fiber has to be routed along streets and rail lines etc., which is generally not a straight line. Finally, fiber cable is implemented with "slack" which allows these to easily be repaired in the case of cable cuts. And all that adds latency.

But there are serious drawbacks. The bandwidth of a microwave link is typically limited and is 1/1000th the capacity of a single fiber pair. This makes the cost per Gigabit very expensive. Microwave dishes require roof access on buildings and microwave towers to cover distances longer than 30 to 40 miles. And finally, microwave is vulnerable to "rain fade," the absorption of the microwave signal by rain and snow, which means that microwave links have much lower reliability than fiber links.

Nabicht: Because of bandwidth limitations, microwave providers can't solely rely on the previous model of dividing bandwidth among enough customers to make building the path worth it. The bandwidth limitations forced providers to come up with other revenue streams.

Market data distribution was an obvious offering. It is an efficient use of the bandwidth available: one copy of the data is sent over the microwave and distributed to multiple

Quincy Data

MICROWAVE FOR THE MAINSTREAM

FAST

Most links are the lowest known latency • All latencies published on www.quincy-data.com

FAIR

All clients have access to the fastest data on equal terms • Three-month contract terms • Discount for Small Firms • All prices published

FLEXIBLE

Purchase only the data channels and locations needed • Illinois – NJ • Chicago Local • 8 New Jersey POPs • London • Other major European Trading Centers coming in 2014 • Choose asset classes from 8 Exchanges

Quincy Data, LLC • 201-884-1114 • 2355 Broadway • Suite 207 • Oakland, CA 94612 • www.quincy-data.com

Inside Market Data

The new bumper issue

Inside Market Data app now available – FREE for all *Inside Market Data, WatersTechnology Data* and *Premium Package* subscribers.

Download all the content from the latest weekly print issue, build your archive of issues and view them offline.

To download and find out more visit
waterstechnology.com/static/imd-app

An incisivemedia website
incisivemedia.com

Inside Market Data Inside Reference Data Buy-Side Technology Sell-Side Technology
waterstechnology

Gil Tene,
Vice President of Technology and
Chief Technical Officer
Azul Systems
Tel: +1 650 230 6555
Email: gil@azulsystems.com
Web: www.azulsystems.com

consumers. The number of customers is only limited by the number of participants in the market who can benefit from the service, so market data distribution was priced aggressively to bring on many clients. The cost of consuming microwave carried market data is much cheaper than the bandwidth needed to transport your own market data on the previous fastest fiber optics paths. Thus microwave has made low-latency market data more widely available and cheaper for companies to obtain.

Yanilmaz: Speed and large bandwidth are obvious advantages. Signal deterioration by objects and atmosphere in the transmission trajectory, and deployment costs are typical disadvantages. Microwave transmission services are expected to expand as alternatives to fiber-optic cable trunks, and continue serving traders for whom the incremental advantages in transmission speed outweigh the microwave's higher cost.

Tene: We don't deal with microwave technologies directly, but our customers use it. Fundamentally, microwaves are about shaving one-third off the latency between physical locations a long way apart. It comes with cost issues and with bandwidth issues, because you can move data more reliably over fiber than through the air.

But microwave is really addressing the antiquated notion that markets are far apart, and trying to make them closer in terms of the time it takes to transmit data between them. Going forward, I see a physical shift taking place in marketplaces, in terms of where they are located—for example, moving multiple exchanges into the same building. I would expect that over time, more venues will start moving their physical venues to cut the time between them—for example, moving matching engines from Chicago to New York to eliminate the latency between Chicago and New York. And the more shared co-location venues we see, the less you will see technologies used to reduce latency over distances. But obviously there are good political and regulatory reasons why exchanges can't come together completely, so these technologies like microwaves will probably be around for a while.

"While the race to being super-fast has made people build very simple trading systems with few parameters, we are already seeing people use tens of gigabytes of memory that are cheaply available today within the critical execution path to apply a lot more real-time information to trades, far beyond what they can capture in just 10 parameters."

Gil Tene, vice president of technology and chief technical officer, Azul Systems

IMD: In the same way that "old" radio waves have been revived to serve new uses, are there other existing, technologies that could be repurposed or repositioned for the latency space? Can any existing latency technologies be used more broadly to achieve better results? Where will the next sources of latency innovation come from?

Tene: It's always hard to guess the next solution, but there are a few things I see people doing today that seem obvious in hindsight. One is the shift from text-based to binary data protocols for trading that are easier for machines to process. It surprises me that we still use text protocols for computer-to-computer trading. Humans aren't reading those messages, but the payloads are still in fat text. You can always translate those messages into text if a human wants to read it. But I get amused when I hear people talking about single-digit microsecond latency figures when most of that delay is spent processing text messages.

We need to get everyone to move to a faster, cheaper network protocol. Some parts of the industry are still using TCP, which has outrageous latency issues because the outlier behavior built into the protocol is very high. So I think we will see movement to other protocols—though I'm not sure which protocol will win out, whether it will be reliable UTP or others—and this shift will have a fundamental effect on latency.

And while the race to being super-fast has made people build very simple trading systems with few parameters, we are already seeing people use tens of gigabytes of memory that are cheaply available today within the critical execution path to apply a lot more real-time information to trades, far beyond what they can capture in just 10 parameters. But this also increases speed, because more memory speeds up those computations, so we are seeing the use of large in-memory computing infrastructures to speed up that critical path of trades.

Ruvo: Latency improvements will likely come in two areas, computation and communications. In communications, the two most interesting technologies are millimeter wave and hollow core fiber. Millimeter wave is much like microwave but allows for more efficient use of the available RF spectrum, which means that more bandwidth can be delivered than across a comparable microwave link. Hollow-core fiber

allows light signals to travel through fiber-optic cable, but in a vacuum in the middle of a fiber rather than through the fiber medium itself. In practice, hollow-core fiber is not yet commercially viable and has severe distance limitations, but it's an interesting space to watch.

In computation, we will see continued progress in latency from several factors: Moore's Law, which allows processors to run faster, although chip makers are getting diminishing returns on things like clock speed; increased use of FPGA and the commoditization of FPGA technology that will make such solutions available to a broader market; use of digital signal processors and GPUs for selective tasks; and the combination of Moore's law and FPGA.

FPGA chips have traditionally gained performance by parallelization of tasks and an architecture designed for data flow, but have not been implemented using the fastest silicon (FPGAs typically have slower clock rates than the latest Intel CPU). Combine faster silicon and FPGA architecture and you will get a lower-latency processor.

Yanilmaz: Hollow-core optical fibers do improve current fiber-optic technology by speeding light transmission through their air-filled cores, but I think this technology requires further development to be practical. Other alternatives such as neutrino transmitters are just experimental, and their potential for commercialization remains to be proven.

Tyc: There will be many innovations inside co-location sites to make latency equal for everybody. This is very important from the point of view of fairness. When you co-locate at the CME facility in Aurora, all the cabling is equidistant. This is very important for trading firms; it means that this is an issue that they do not have to worry about. Everybody will have the same distance to the matching engines and firms can focus on matters more related to actual trading.

Peter Nabicht
Modern Markets Initiative

Nabicht: Each major latency-reducing technology seems to have fewer and fewer customers because the investments are substantial and firms only make the investment if the latency improvements make economic sense for the strategies they are running.

I'm not convinced the next latency improvements come from technology. I think they come from quantitative research and strategy improvements. If you can make a decision earlier on a different or smaller sub set of data—not having to wait for as much data in order to determine what your action should be—then in effect you've

gained a latency improvement. For example, take a string of 20 updates. If my strategy currently makes a decision on the twentieth update but I can find a way for it be as profitable or almost as profitable acting on the tenth update then I just got faster by however much time it takes for update eleventh to the twentieth update to come.

IMD: What is the future of the latency race? Will ultra-low-latency become a priority again when the market becomes more bullish, or is latency becoming over-commoditized—at least in equities? How else can firms that can no longer compete on speed achieve a competitive advantage?

Tyc: The obvious future of the latency race is delivering markets at the speed of light. Quincy is barreling down that path and we hope to get there first. For risk takers and managers, the important question is who they can rely on to consistently deliver—in terms of latency, reliability and data breadth—and what do they need to run their businesses effectively today and at each step toward zero latency? We have a very strong track record of delivering for clients. In the end, firms no longer need to compete on speed, they need to partner on speed. Our goal is to be the partner of choice.

Yanilmaz: Latency itself isn't becoming commoditized; the means of data transmission are. Effective real-time tracking and management of latency as a key, multi-dimensional statistical metric across trading firms' whole ecosystem will remain a key proprietary and competitive advantage.

Mehmet Yanilmaz
Myra Trading

Ruvo: Ultra-low latency remains a priority for firms with the HFT strategies that can exploit it and the scale to justify the expenditure. But it's not for everyone. Most firms never did compete on speed. Some compete with sophisticated computer based strategies. Others study the fundamentals of a business, meet with management, walk the floors of factories and make long-term bets. There are many ways to make money and while latency matters to some of those strategies, it was never for everyone.

Nabicht: Low latency might be necessary to maximize some strategies, but it is never sufficient [on its own]. Low latency has never been the all-important competitive advantage. It is a tool that a strategy can use, but its importance always ranks behind pricing and risk management. It doesn't matter how fast you are if your price is wrong; it doesn't matter how fast you are if you can't properly manage your risk. There are many strategies that do not need to compete on speed—they compete on smarts.

As low latency becomes more and more commoditized and the differences between the leaders and the pack decreases, more companies will be more competitive across a larger array of strategies. The competitive focus will be where it has always has been—on the strategy. The ability to price the market and manage the risk of your strategy has always been important; the difference going forward is that we'll recognize this more and talk less about low latency.

Tene: Whenever I hear people talk about getting out of the latency race, it's always a decision to divest because they are not making money anymore. The people who are winning because they have the best latency behavior are not backing off; only the losers will back off. And just because a firm says there is no money to be made from chasing latency doesn't mean that's correct—it just means that they can't make money from it, so they are cutting their losses.

But others are finding ways to use low-latency technologies in more profitable ways. For example, equities have operated at low latency for a long time, and after a point, it becomes a game of diminishing returns. But other asset classes are ripe for the benefits of low latency, so if we look at the history of latency in the equities markets, we should expect that to repeat itself in other asset classes. And if you've built a good practice in one asset class, you can do it in others: the technology and skills are transferable, and there is a lot of unpicked fruit in other fields.

Take foreign exchange, for example, which operates at a couple of orders of magnitude higher than equities. So if you squeeze out latency in a smart way, you can make room to compete in FX. I don't think anyone is easing up on the latency race in FX.

And in other asset classes, more volume is moving to electronic trading, indicating that latency will become more and more important as the shift to e-trading improves latency. Then once there is enough liquidity, we will see more venues form, and people will start to game latency in ways that they haven't before.

This is where latency volatility becomes particularly interesting, and is not just about risk, but also about opportunity. If it takes me three milliseconds to execute a trade and the market moves quicker, that could have a significant impact on my position. But others will seek to benefit from that volatility by watching you and understanding your latency volatility. And if they know your system sometimes stalls for five milliseconds, and they see you stalled for one millisecond, then they know you're going to be stalled for four more milliseconds, and can trade against you.

People understand this, but there is a lack of attention to the behavior of latency and how it can improve systems to predict reliability or how to exploit it to create opportunities. Latency isn't just a number; it's a behavior—and there is a lot of opportunity in understanding this. ■

SPONSOR'S STATEMENT

Latency: Dead or Just Misunderstood?

Is the race for latency dead, or is low latency something that everyone needs to have? Both of these claims have been made, but both are overly simplistic. To understand why latency matters, we first need to understand a fund's investment strategy and therefore its trading strategy, says Bill Ruvo, global business head of real-time feeds, Enterprise capabilities at Thomson Reuters.

If we consider that financial markets are efficient over some period of time, then eventually all information about a security will be "priced in." In the long term (on human scale), this could mean that the success of a new product launch is eventually realized in the form of increased profitability, resulting in a rise in the share price. Over shorter periods of time, information comes in the form of changes to supply and demand in financial markets, which translates to shorter-term price movements until a stock reaches a short-term equilibrium. This process may be carried out by people looking at screens and reading news, or by computers reading direct exchange feeds and running predictive models. Either way, an investment strategy begins with the receipt of information, analysis of that information to determine a projected view of the future price of the security, and execution of a trade to capture the difference between the current price and the future price.

Information "signals" span a broad range of time sensitivity, levels of certainty and potential price impact. Different fund managers with disparate interests, skills and mandates from their investors will naturally focus on different types of information and investment strategies. Those strategies will vary in their information needs and therefore in their sensitivity to latency.

At one end of the spectrum are high-frequency traders (HFTs) with simple strategies and a focus on signals that are very short term in nature. They tend to have high portfolio turnover, short holding times and a very high degree of latency sensitivity, because the signals consumed by their strategies decay very quickly. As strategies become more complex, typically with broader information and compute requirements, they are able to identify trading opportunities that would be missed by simpler strategies and where the alpha decay is longer, so they are less latency sensitive.

At the extreme opposite end from HFT funds are classic, fundamental "Graham and Dodd" style investors. The signals they consume have little to do with real-time market data and much more to do with understanding the management of a company, its strategy, market position and broader economic factors that affect its market. These investors care very little about latency, but they do care a great deal about depth of information and quality of analysis. Their holding periods are long and their portfolio turnover is very low.

There are, of course, many funds that operate all along this latency spectrum, which isn't readily divided into discrete segments. Most fund managers, with the exception of the very smallest, follow a range of strategies. They will shift their focus based on the potential trading profits and the costs of participating in those strategies.

So does latency still matter? To paraphrase Yogi Berra, nobody trades HFT anymore, it's too crowded. Of course latency matters. However, over the past four to five years the volumes and volatility in many markets (which create opportunity for HFTs) have declined, and the cost to participate (buying/building proprietary FPGA systems, leasing expensive co-location and communications lines) has increased. This has resulted in consolidation of the ultra-latency sensitive end of the market, and even traditional ultra-HFT firms have diversified into strategies that are less latency sensitive and more computationally/data complex.

The challenge for the vendor community is threefold: (a) to deftly and nimbly meet changing market requirements; (b) to deliver value as these requirements shift; and (c) to remain relevant as workflows evolve. How is this accomplished? In short, by offering a suite of interoperable propositions with varying latency characteristics, appropriate for different use cases and that effect minimal impact on a customer as his requirements change or expand. Not only does speed apply to the delivery and ingestion of market data and related information for most trading strategies, but speed is also important in terms of how quickly a firm can shift to, or adopt, new strategies.

Based on market activity and new regulation, a firm's success can be attributable to its ability to develop and roll-out new strategies and this is either facilitated or hindered by how quickly new strategies can be back-tested and developed. Development cycles must be compressed without compromising the quality of inputs. And as noted above, alpha generation for high-frequency strategies is particularly short lived, so ease of development and the ability to move up—and down—a familiar proposition stack (consistent data models, APIs, symbology) can mean the difference between success and failure for a firm facing volatile markets and changing regulation.

So while latency is still more than relevant, ease and speed of development (regardless of the strategy's latency requirements) can be equally important to a firm's ultimate success. ■

Inside Market Data | Inside Reference Data | Buy-Side Technology | Sell-Side Technology

waterstechnology

2014 events

Hosted by Inside Market Data, Inside Reference Data, Buy-Side Technology, Sell-Side Technology and Waters magazine, the WatersTechnology series of events are the leading financial data management and technology conferences for information and systems professionals working at financial trading firms around the world.

These conferences deliver expert analysis and commentary through interactive panel discussions, case studies and keynote addresses that provide delegates with the latest on the business, competitive, regulatory and technological issues affecting market data, reference data and trading technology professionals.

Our series of events provide the opportunity to network with hundreds of leading market data, reference data and trading technology executives from consumers, producers and suppliers across North America, Europe and Asia Pacific.

April

North American Trading Architecture Summit 2014

Tokyo Financial Information Summit 2014
Inside Market Data
Inside Reference Data

Tokyo Trading Architecture Summit 2014

May

Buy-Side Technology European Summit 2014

North America Financial Information Summit 2014
Inside Market Data
Inside Reference Data

June

Buy-Side Technology Asian Summit 2014

Asis Pacific Trading Architecture Summit 2014

Toronto Financial Information Summit 2014
Inside Market Data
Inside Reference Data

June

Toronto Trading Architecture Summit 2014

September

European Financial Information Summit 2014
Inside Market Data
Inside Reference Data

October

Frankfurt Financial Information Summit 2014
Inside Market Data
Inside Reference Data

Buy-Side Technology North American Summit 2014

November

European Trading Architecture Summit 2014

Asia Pacific Financial Information Conference

December

waters USA 2014

For more details about sponsoring or exhibiting contact:
Jo Webb
T: +44 (0)20 7316 9474
E: jo.webb@incisivemedia.com

To register as a delegate for one of our events contact:
Sam Lawson
T: +1 (646) 755 7399
E: sam.lawson@incisivemedia.com

An incisivemedia event
incisive-events.com

REUTERS/Claro Cortes

Thomson Reuters Elektron delivers low latency feeds, along with the analytics, platform and transactional connectivity to support any financial workflow application – including Thomson Reuters Eikon. All capabilities can be deployed at your site, or delivered as a fully managed service.

financial.thomsonreuters.com/elektron

THOMSON REUTERS™


An
incisivemedia
publication